*Computed and actual yields of apples in New York State*

[Yields in millions of barrels]

| Year | Computed yield | Actual yield | Difference |
|---|---|---|---|
| 1901 | 6.0 | 3.7 | 2.3 |
| 1902 | 12.1 | 13.7 | 1.6 |
| 1903 | 16.2 | 15.3 | 0.9 |
| 1904 | 12.0 | 18.3 | 6.3 |
| 1905 | 8.5 | 7.0 | 1.5 |
| 1906 | 9.9 | 10.3 | 0.4 |
| 1907 | 8.3 | 9.3 | 1.0 |
| 1908 | 11.6 | 11.0 | 0.6 |
| 1909 | 7.5 | 8.5 | 1.0 |
| 1910 | 5.7 | 5.7 | 0.0 |
| 1911 | 14.6 | 13.0 | 1.6 |
| 1912 | 11.6 | 14.7 | 3.1 |
| 1913 | 6.8 | 6.5 | 0.3 |
| 1914 | 15.0 | 16.5 | 1.5 |
| 1915 | 6.6 | 8.5 | 1.9 |
| 1916 | 12.0 | 11.8 | 0.2 |
| 1917 | 11.0 | 5.4 | 5.6 |
| 1918 | 17.2 | 13.6 | 3.6 |
| 1919 | 6.7 | 4.8 | 1.9 |
| 1920 | 14.1 | 15.7 | 1.6 |
| 1921 | 6.3 | 4.5 | 1.8 |
| 1922 | 10.8 | 12.0 | 1.2 |
| 1923 | 8.0 | 8.3 | 0.3 |
| 1924 | 8.1 | 7.3 | 0.8 |
| 1925 | 7.9 | 10.8 | 2.9 |
| Average | 10.2 | 10.2 | 1.8 |

It will be seen that there are several instances where the computed yields show large deviations from the true yield, but these are not as large as their deviation from the average yield. The standard deviation of yield is 4.05 million barrels and that of actual from computed is 2.33 million barrels, or a reduction of 42.5 per cent.

## SUMMARY

The data on hand are, of course, rather limited and can not take into account all possible influences on yield. It was planned originally to demonstrate that apple yields were largely affected by spring temperatures and this seems to be proven beyond a reasonable doubt.

There are, of course, other factors which influence yield, but in a study of this type for an entire State they are too varied to be included and an attempt to combine all possible influences, if known, would necessarily be tremendously bulky and take an amount of time entirely out of comparison with the results obtained.

Single orchards, if complete data could be obtained, would produce results of more significance than those for a whole State. The State data must necessarily be less complete and more difficult of access even when there are more or less detailed reports. Using the data before mentioned the results are very satisfactory in that they conclusively demonstrate that the one factor of major importance is spring temperatures.

### LITERATURE CITED

(1) Livingston, Burton E.
1916. PHYSIOLOGICAL TEMPERATURE INDICES FOR THE STUDY OF PLANT GROWTH IN RELATION TO CLIMATIC CONDITIONS. Physiological Researches, vol. 1, 8: 399–420.

(2) Seeley, D. A.
1917. RELATION BETWEEN TEMPERATURE AND CROPS. Mo. Wea. Rev., **45**, 7: 354–359.

(3) Wallace, H. A., and Snedecor, George W.
1925. CORRELATION AND MACHINE CALCULATION. Official Publication, Iowa State College, **23**: No. 35.

# ON THE MEASURE OF CORRELATION

By Gilbert T. Walker

[Imperial College of Science and Technology, South Kensington, London, S. W. 7, November 1, 1927]

There has of late been a welcome recognition of the services that can be rendered to meteorology by statistical methods; but associated with some of the recent theoretical discussion there have been elements which appear to me unsound and I would ask permission to make some remarks on a theorem which is attributed to W. H. Dines.

1. The authoritative enunciation of the theorem is that contained in the Méteorological Magazine.[1]

"If there is a cause $A$ and a result $M$ with a correlation $r$ between them, then in the long run $A$ is responsible for $r^2$ of the variation of $M$."

On the other hand, working in India in regrettable ignorance of the classical literature of the subject, I was led to develop the ordinary regression equations from a definition of the correlation coefficient between two quantities as "the proportionate extent to which the variations of each are determined by, or related to, those of the other." [2]

2. It might at first sight appear that so fundamental a discrepancy must rest on a wide difference of terminology; but this can scarcely be the case. If the departures of $M$ and of $A$ are denoted by $x_0$ and $x_1$, and their standard deviations or "square-means" by $\sigma_0$ and $\sigma_1$, we may denote $x_0/\sigma_0$ and $x_1/\sigma_1$, "the proportional departures," by $z_0$ and $z_1$.

Then the ordinary regression equation is

$$x_0 = \frac{r\sigma_0}{\sigma_1}x_1 + b$$

where $b$ is independent of $x_1$, or $z_0 = rz_1 + d$, where $d$ is independent of $z_1$.

That part of the variation of $M$ which is related to, or controlled by, $A$ is, by (1), $r\ \sigma_0\ x_1/\sigma_1$; and it is important to note that this value is accepted by both parties in this discussion. In the last paragraph of the statement of the Meteorological Magazine we read "the average contribution of $a$ to $m$, i. e., the average value of $r\sigma_m\left[r\frac{m}{\sigma_m} + y\right]$"; and, by equation (7) there, this is equal to $r\sigma_m\left[\frac{a}{\sigma_m}\right]$; in our notation this is $r\sigma_0\frac{x_1}{\sigma_1}$, which bears to $\sigma_0$ the ratio $rz_1$. We may note that this interpretation is also accepted by Krichewsky,[3] who writes in his (6a) the regression equation for two variables as $\bar{z}_0 = \beta_{01}z_1$ and replaces this in his (11) by $\bar{z}_0 = r_{01}\beta_{01}z_1$. He then defines $E_{01}$ as "that part of the variation of $z_0$ for which the variable $z_1$ is responsible in the long run . . " and takes $E_{01}$ as $r_{01}\beta_{01}$.

, Now, as stated below, I do not agree with the substitution of $\beta_{01}z_0$ for $z_1$, but the fact remains that Krichewsky regards something equal to $\beta_{01}z_1$ as the part for which $z_1$ is responsible.

3. Now $x_1$ is a quantity obeying the same error law of distribution as $x_0$, its standard derivation being $\sigma_1$ corresponding to $\sigma_0$ for $x_0$; so just as the values of $z_0$ obey the error law of distribution and have a standard deviation of unity, the values of $rz_1$ will obey the error law and have a standard deviation of $r$. To say that in the long run these values of $rz_1$ are $r^2$ times those of $z_0$ appears to me definitely because mathematically, incorrect. It must

[1] February,1921, p. 21.

[2] Indian Meteorological Memoirs, Vol. xx, Pt. 6, p. 120, 1909.

[3] "Interpretation of correlation coefficients." Physical Dept. Paper No. 22, Cairo, 1927.

be admitted as conceivable that on general grounds a man may prefer to estimate the figure of merit of a correlation as measured by $r^2$ and not by $r$; but this does not give him the right to say that if the terms of one group of figures are $r$ times those of another, the ratio of one group to the other is $r^2$.

4. The error creeps in when Krichewsky replaces $z_1$ by $\beta_{01}z_0$ or $rz_0$; for when forecasting it is $z_1$ that is given and the estimated value of $z_0$ is $z_0+e$, the error being independent of $z_1$. But the mean value of $z_1$ would be $rz_0$ if we were forecasting $z_0$ from $z_1$ by an equation $z_1=rz_0+f$, and the error $f$ in that forecast would be independent of $z_0$, which is quite a different matter. If it were legitimate to replace a quantity by its mean value under different conditions we could apparently carry the process further and derive the impossible equation $z_0=rz_1=r^2z_0=r^3z=r^4z_0= \cdot \cdot \cdot$

But if we replace $z_1$ by $rz_0+f$, to which it is equal, and note that the standard deviation of $f$ is $(1-r^2)^{1/2}$, we see that the standard deviation of $r\ (rz_0+f)$ is $r$ times the standard deviation of $(r^2z^2_0+f^2)^{1/2}$, or $r\ \ (r^2+1-r^2)^{1/2}$, which is $r$ not $r^2$.

A further point is that the proof of the $r$ law just given holds whether or not there are other factors not independent of $z_1$.

5. The only argument with which I am acquainted for wishing to estimate relationships by $r^2$ rather than $r$ is that if a quantity were controlled by two independent factors the total relationship would then be got by adding the component relationships. To this the reply is that in meteorology independence is the exception not the rule. If pairs of forces acting on a particle were always at right angles it might in the same way be urged that the effect of a force should be estimated by its square in order that the resultant might be estimated by the sum of the forces. Now, in estimating the value of a method of forecasting the proportion to which the forecast is controlled by the known data is in my opinion the vital feature, and I should not regard it as more justifiable to adopt $r^2$ rather than $r$ because it would have points of convenience in exceptional cases than I should to measure forces by the squares of their present measures for a similar exceptional convenience.

## NOTE ON THE THEOREMS OF DINES AND WALKER

By EDGAR W. WOOLARD

Let $x_0$, $x_1$, be the departures of any two varying quantities; and let the (unknown) complete and exact functional relation in which they are involved be

$$F(x_0, x_1, x_2, \cdot \cdot \cdot \cdot \cdot \cdot \cdot)=0, \qquad (1)$$

in which $F$ may be of any form, and in which the $x_i$ may be mutually dependent in any manner, or in part mutually independent.

From a number of pairs of corresponding observed values, we may always compute $\sigma_0$, $\sigma_1$, and $r$. Furthermore, for any individual pair we can always write

$$\frac{x_0}{\sigma_0}=r\frac{x_1}{\sigma_1}+b, \qquad (2)$$

because a value can always be assigned to $b$ so that this equality will be satisfied; similarly we can always write

$$\frac{x_1}{\sigma_1}=r\frac{x_0}{\sigma_0}+b'. \qquad (3)$$

Also, for any *given fixed* value of $x_1$, we can always find $B$ such that

$$\frac{\bar{x}_0}{\sigma_0}=r\frac{(x_1)}{\sigma_1}+B, \qquad (4)$$

and for any *given fixed* $x_0$ we can find $B'$ such that

$$\frac{\bar{x}_1}{\sigma_1}=r\frac{(x_0)}{\sigma_0}+B', \qquad (5)$$

in which $\bar{x}_0$, $\bar{x}_1$, are the *means* of the values of one variable associated with a *fixed* value of the other. The curves

$$x_0=r\,\frac{\sigma_0}{\sigma_1}\,x_1,\ x_1=r\,\frac{\sigma_1}{\sigma_0}\,x_0, \qquad (6)$$

are the straight lines of "best fit" (in the sense of least squares) to the individual observations and to the means. However, the fit may or may not be close, and in either case there may or may not exist systematic departures from it; $b$, $B$, may or may not be independent of $x_1$, e. g., and certainly will not if $x_1$ is not independent of $x_2, \cdot \cdot \cdot \cdot \cdot \cdot$ The standard deviations of $b$, $b'$, are each $(1-r^2)^{\frac{1}{2}}$.

The preceding equations do not, by themselves, permit any conclusions whatever to be drawn concerning relations of cause and effect; they apply to mere covariation only.

Sir Gilbert Walker has defined the correlation coefficient $r$ as "the proportionate extent to which the variations of each of two quantities are determined by, or related to, those of the other," whence "if there is a cause $A$ and a result $M$ with a correlation $r$ between them, then in the long run $A$ is responsible for a fraction $r$ of the variations of $M$." The exact meaning intended to be conveyed by this statement is to be found in the mathematical reasoning by which the theorem is supported:

*If*, in (2), $b$ *is* independent of $x_1$, then the part of the variation of $x_0$ which is controlled by $x_1$ is $r\,\frac{\sigma_0}{\sigma_1}\,x_1$, and the standard deviation ("square mean") of this controlled part is $r$ times the standard deviation, or mean variation of $x_0$. From this it appears that Walker adopts the standard deviation as a measure of variation and intends his theorem to state that a fraction $r\sigma_0$ of $\sigma_0$ is due to variations in $x_1$, and the remainder $(1-r)\sigma_0$ to variations in $x_2, \cdot \cdot \cdot \cdot \cdot \cdot$ Clearly, this implies not only that $x_1$ is independent of the remaining variables, but also that $x_0$ and $x_1$ are linearly related, so that $b$ is a function only of $x_2, \cdot \cdot \cdot \cdot$ ; in this case, the first term on the right of the identity

$$\sigma_0{}^2=r^2\sigma_0{}^2+(1-r^2)\sigma_0{}^2 \qquad (7)$$

is, by (2), the fraction of $\sigma_0{}^2$ due to $x_1$.

Now, Dines's theorem states that "if there is a cause $A$ and a result $M$ with a correlation $r$ between them, then in the long run $A$ is responsible for $r^2$ of the variation in $M$." Again, the exact meaning intended must be sought in the mathematical proof offered for the theorem:

Substitute (3) in (2):

$$x_0=r\sigma_0\left[r\frac{x_0}{\sigma_0}+b'\right]+b\sigma_0. \qquad (8)$$